Web Application Security and Search Engines – Beyond Google Hacking

Amichai Shulman, CTO, Imperva Inc.
Agenda

- Google Hacking on Steroids
  - Automated Google Hacking
  - Google Worms
- Malware Distribution & Search Engines
- Site Masking
- The Search of Death
- Google for Security
In the News
Accidental Data Leakage

**STONY BROOK INDEPENDENT**

**Personal Information of up to 90,000 Compromised at Stony Brook**

The personal information of 90,000 people in a Stony Brook University database was accidentally posted to Google last week until it was discovered almost two weeks later.

The letter indicated the files were not easily accessible through Google and that the “information could only be retrieved through the use of multiple criteria.” It said the New York State Cyber Security Office contacted Google to have the information removed after it was discovered on April 24.

---

**Coastal**

Last modified Thursday, July 12, 2007 10:31 PM PDT

**Confidential data revealed on Encinitas’ Web site**

ENCINITAS -- Credit card or checking account information and addresses for nearly 1,200 people who had enrolled in Encinitas’ youth recreation programs was inadvertently posted on the city’s Web site, officials said Thursday.

---

**COMPUTERWORLD UK**

May 18, 2007

**Minister pledges inquiry into UK visa website security breach**

How in VFS site allows access to personal information on Visa applications

---

**NU Contacting 4,000 After Security Breach**

Students And Applicants’ Personal Information Became Available Online, School Says

© Jun 1, 2007 9:55 pm US/Central
In the News
Accidental Data Leakage

Patient data exposed online
Error led HMO to post information on 75,000 in Md., D.C.
By Liz F. Kow | Sun reporter
March 26, 2006

A CareFirst BlueCross BlueShield dental HMO accidentally exposed personal information, including Social Security numbers, of about 75,000 members on a public Web site last month and didn't notify them until about three weeks later.

The Dental Network, which is owned by CareFirst, informed the members - mostly Maryland and District of Columbia residents - that their names, addresses, dates of birth and Social Security numbers had been posted on its Web site for two weeks in February because of a technical error.

The company says that to its knowledge, no one has misused the information "the risk should be taken seriously" and has offered members 12 months of identity monitoring, as well as information about contacting the three credit fraud alert on their account.

Pennsylvania Yanks Voter Site After Data Leak
Robert McMillan, IDG News Service
Wednesday, March 19, 2003 6:20 AM PDT

Recommended this story? Yes 8 Votes No 0 Votes

With voting in Pennsylvania's presidential primary just a month away, the state was forced to pull the plug on a voter
Google Hacking on Steroids

- What is Google Hacking?
  - Using a search engine to uncover application vulnerabilities or sensitive data
  - Most notable resource is Johnny Long’s Google Hacking Database

- What is the threat to application owners?
  - All Web site content is exposed to Google
  - Sensitive content might be available for months before the compromise is discovered
  - Even after sensitive pages are removed, they will be stored in Google Cache
Automated Google Hacking

- Automating data leakage discovery and vulnerability discovery
- Powerful hacking tool – quickly find a list of vulnerable sites per set of vulnerabilities

Tools
- Goolag Scan
- Gooscan
Automated Google Hacking

- Google are putting a strict restriction on the number of queries per IP per day.
  - Violating IP addresses are punished by having to answer a Captcha
- Hackers are mainly unaffected
  - Can use a multitude of open proxies on the web
  - Need only scan for a small set of vulnerabilities
Google Worms

- Concept first analyzed in March 2004 ADC paper: “Web Application Worms: Myth or Reality?”
- First actual exploit - Santy Worm, December 2004
- What is a Google Worm
  - Normal worms randomly draw IP addresses and hope for the best.
    - Large foot print
    - Infection is almost impossible for non-standard deployments
  - Google Worms search Google for the first batch of vulnerable sites, infect them and instruct the infected machines to go for a different batch each.
    - Small foot print
    - Infection guaranteed regardless of deployment differences
Google Worms

Recent Incidents

January 2008

- First massive incident combining SQL Injection and HTML Injection
- Tens of thousands of sites affected (including CA’s site)

April 2008

- Huge number of MS SQL Server based sites injected with HTML IFRAME through SQL Injection
- Infection code introduces an IFRAME to each field in the database.
- Number of infected hosts and their uniformity suggest an infection engine based on Google Hacking

May 2008!
Mitigation Strategies

Google Hacking

- Passively monitor Web traffic for sensitive information
  - Preventing data leakage without affect application delivery is a difficult problem. Preventing leakage of information to search engines is much easier
  - Search request / reply pairs for potentially sensitive information. Block reply if request is made by a search bot.
  - Need a solution that contains an up-to-date database of search terms equivalent to those found on Johny Long’s site. Additionally, a security policy should correlate a match to one of these terms with search engine IP addresses and User-Agent headers
Mitigation Strategies

Google Hacking

- Actively Search Google (or any other search engine) for leakage
  - Difficult to use a tool like GoolagScan. Google’s anti-automation measures do affect application owners trying to defend themselves.
  - Requires a custom engine with relatively slow rate, and a database that is always up-to-date with latest Google Dorks.
Malware Distribution & Search Engines

- How can search engines be used to distribute malware, or other attack vectors?
  - Infect a page on the web
  - Make sure that the page is ranked high for popular search terms

- Google study concluded that about 1.3% of search queries returned at least one malicious URL result

- How can I ensure that the infected page is ranked high for popular search terms?
  - Aha!
Malware Distribution & Search Engines

www.famouse.com

search.asp?query="Paris Hilton" + xss vector

Attacker

Victim

Google
Malware Distribution & Search Engines

www.famous.com

search.asp?query="Paris Hilton" + xss vector

Attacker

Victim

Google
Malware Distribution & Search Engines

www.famous.com

search.asp?query=“Paris Hilton” + xss vector

Index site

Google

Attacker

Victim
Malware Distribution & Search Engines

www.famous.com

search.asp?query="Paris Hilton" + xss vector

Index site

Google

Attacker

Victim

“Paris Hilton”
Malware Distribution & Search Engines

www.famouse.com

search.asp?query="Paris Hilton" + xss vector

Index site

Google

"Paris Hilton"

Attacker

Victim
The example is taken from a recent attack that took place on March 2008 and described by researcher Dancho Danchev.

Results from major web sites rank high in search engines.

Many sites allow search engines to index pages with internal search results.

An attacker found XSS vulnerabilities in search pages of high profile sites. Attacker then used the search functionality to look for popular search terms (e.g. Paris Hilton), appending the attack vector as part of the search.

The result pages with the attack vector embedded in them are then ranked high for those popular search terms and Bingo!
Mitigation Strategies

Malware Distribution

- Careful input validation and sanitation is always a good practice
- Fast reaction using up-to-date signature mechanism can provide timely protection against a sudden outbreak of an attack and ensure that malicious content isn’t delivered by application to users
- Search engines are trying to identify “infected” pages and place a visual notification
  - Yahoo! SearchScan
  - Google ("This site may harm your computer")
Mitigation Strategies

Malware Distribution
Site Masking

- **What is it?**
  - Take your competitor out of Google!
  - Index a competitor’s content under your domain!

- **How?**
  - Can only be applied to relatively small scale sites
  - Google penalizes sites for having duplicate content
  - Make Google believe that the original content is actually a copy
Site Masking

BBC NEWS | Technology | Hackers hijack web search results - Mozilla Firefox

Last Updated: Thursday, 29 November 2007, 11:31 GMT
Email this to a friend |  Printable version

Al Gore's Blog Hacked

by Tim Stevens, posted Nov 27th 2007 at 9:59AM

The booby-trapped sites have been removed from Google's search results.

A huge campaign to poison web searches and lure people into visiting malicious websites has been thwarted.

The booby-trapped websites came up in search results for search terms such as 'Christmas gifts' and 'hospital drugs'.

Windows users falling for the trick risked having their computer hijacked and personal information plundered.

The criminals poisoned search results using thousand of domains set up to convince search-index software...
Site Masking

Google Bot

Malicious Web Page
www.mal.com

Proxy
www.proxy.com

Original Web Page
www.foo.com
Site Masking

1. Malicious Web Page
   www.mal.com

Google Bot

Proxy
www.proxy.com

Original Web Page
www.foo.com
Site Masking

Google Bot

Malicious Web Page
www.mal.com

Proxy
www.proxy.com

Original Web Page
www.foo.com

... href=http://www.proxy.com?url=www.foo.com
Site Masking

1. Malicious Web Page
   www.mal.com

2. Original Web Page
   www.foo.com

3. Proxy
   www.proxy.com

... href=http://www.proxy.com?url=www.foo.com
Site Masking

1. Malicious Web Page
   www.mal.com

2. Google Bot

3. Proxy
   www.proxy.com

4. Original Web Page
   www.foo.com

... href=http://www.proxy.com?url=www.foo.com
Site Masking

Google Bot

Malicious Web Page
www.mal.com

Proxy
www.proxy.com

Original Web Page
www.foo.com

... href=http://www.proxy.com?url=www.foo.com
Site Masking

Google Bot

Malicious Web Page
www.mal.com

Proxy
www.proxy.com

Original Web Page
www.foo.com

... href=http://www.proxy.com?url=www.foo.com
Site Masking

- In some cases, by creating many proxy links, Google can be confused to consider the original web site as presenting duplicate content.
- The original web site vanishes from the search result.
Mitigation Strategies

Site Masking

- Based on the source of the request make changes to the outgoing HTML document:
  - If the request is not from a validated robot (user agent header and IP address) then add a noindex in the Robots <META> tag of the page.
  - As a consequence GoogleBot will only index a page if it is accessed directly and not via a proxy.
When Google is an attacker’s weapon of choice…

- Google can access sites that are not open for anonymous public access
- Attack cannot be linked to the source

How do I do that?

- A number of methods under research
- Google can be instructed to follow a link that contains an attack vector
Google for Security – Application Owner

- Google Webmaster Tools
  - Who is searching my site and for what?
- Stop Google Worms Outbreak
- Remove Infected Pages from Results
Google’s Safe Browsing API enables client applications to check URLs against Google's constantly updated blacklists of suspected phishing and malware pages.

GooDelete tool can be used to clear cached Google Toolbar queries that may contain sensitive information that you don't want lying around.

Google will notify user if link is suspicious (see previous slides)
Thank You

Imperva, Inc.

3400 Bridge Parkway, Suite 101, Redwood Shores CA 94065
Sales: +1-866-926-4678    www.imperva.com